I Wrote This Because of That: Interpretable Radiology Report Generation via Visual Prototypes

Marco Salmè[1], **Federico Siciliano**[2], **Fabrizio Silvestri**[3], **Paolo Soda**[1,4], **Rosa Sicilia**[5,†], **Valerio Guarrasi**[5,†]

[1]Campus Bio-Medico University of Rome, [2]Istituto Nazionale di Geofisica e Vulcanologia, [3]Sapienza University of Rome, [4]Umeå University, [5]UniCamillus-Saint Camillus International University of Health Sciences

Abstract

Automated Radiology Report Generation (RRG) could substantially reduce documentation workload by supporting physicians in routine clinical practice. Clinical adoption, however, requires each generated finding to be supported by inspectable evidence that a physician can verify. Existing systems either provide only post-hoc explanations weakly coupled to the prediction, or require costly region-level supervision to achieve spatial grounding. To address this gap, we introduce *ProtoRRG*, an RRG framework that grounds generation in visual prototypes. Specifically, each generated finding is tied to a dual form of evidence: localised image exemplars learned without anatomical supervision, and retrieved prior reports that supply case-based linguistic precedents. A multi-task training strategy reconciles the competing representational demands of prototype-based prediction and autoregressive decoding within a shared architecture. On MIMIC-CXR, *ProtoRRG* improves state-of-the-art clinical F1 and recall by 2.9% and 10%, respectively, with gains also demonstrated on IU X-Ray. An evaluation against anatomical annotations confirms that prototype activations localise within the correct regions, indicating spatially faithful evidence for the generated findings. These results demonstrate that combining visual prototypes with case-based retrieval improves both diagnostic accuracy and report quality while preserving the inspectable evidence that clinical deployment demands. The project page is available at: https://github.com/arco-group/protoRRG

1 Introduction

Automated Radiology Report Generation (RRG) has advanced rapidly, with recent systems producing fluent, clinically plausible reports from medical images (Sloan et al., 2024). Fluency and plausibility, however, are not sufficient for clinical trust: clinicians must be able to trace each generated finding to specific image evidence. Several strategies have been explored to improve interpretability in RRG, but most provide post-hoc explanations that may not reflect the computations actually driving generation (Wang et al., 2024b; Chen et al., 2024). Inherently interpretable alternatives, such as concept-based bottlenecks (Alam et al., 2025; Salmè et al., 2026) and region-guided architectures (Tanida et al., 2023; Bannur et al., 2024), instead require costly additional supervision that is often unavailable. Prototype-based networks (Chen et al., 2019; Nauta et al., 2023) address both limitations: each prediction is derived by assessing whether input regions *look like* learned exemplars, producing localised visual evidence that can be directly verified on the image, without additional annotation. In medical imaging, prototypes have proven effective for classification (Barnett et al., 2021; Kim et al., 2021), but they have not been used as the visual evidence linking each reported finding to its image support in a generative setting.

Extending visual prototypes to RRG raises a central technical challenge: prototype predictors require features that are discriminative and tightly clustered around pathological patterns, while autoregressive decoders rely on rich spatial context to generate complete reports. An interpretable report generator must therefore reconcile both demands within a shared representation. Moreover, localised visual evidence identifies *where* a finding originates but does not determine *how* to express it in clinical language. Retrieval-Augmented Generation (Endo et al., 2021; Song et al., 2024) fills this gap: reports from similar cases supply linguistic precedents and case-based context that clinicians can examine alongside the image evidence.

To this end, we introduce *ProtoRRG*, an interpretable-by-design RRG model that grounds generation in two complementary forms of inspectable evidence: localised prototype activations

[†]Equal Contribution.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 36511–36527
October 24-29, 2026 ©2026 Association for Computational Linguistics

in the image, and retrieved similar reports that provide case-based context for the output.

Our contributions are as follows:

- We introduce, to our knowledge, the first RRG framework where visual prototypes serve as localised, inspectable evidence for each generated finding.
- We propose a shared spatial representation that reconciles discrete prototype activations with dense visual features, resolving the core tension between prototype-style interpretability and autoregressive language decoding.
- We show that visual prototypes and retrieval-based textual cues combine naturally within a unified architecture, yielding competitive report quality on two publicly available datasets, while exposing inspectable evidence for each prediction.

2 Related Work

Radiology Report Generation. Automatic RRG aims to produce clinically accurate free-text reports from medical images. Early approaches cast RRG as an image-to-text generation problem, where visual features are encoded and decoded into reports using transformer-based architectures (Chen et al., 2020; Cornia et al., 2020; Nicolson et al., 2023; Wang et al., 2023; Sun et al., 2025). These methods improve language fluency, but the correspondence between individual generated findings and the image evidence supporting them is generally not made explicit. As a result, they may miss subtle abnormalities or generate clinically plausible statements that are weakly supported by the image. To bridge the modality gap between radiographs and reports (Liang et al., 2022), several methods adopt vision-language pretraining to learn stronger image-text alignments (Yan and Pei, 2022; Li et al., 2023b; Mei et al., 2025; Wang et al., 2025b). Although these methods improve both text quality and clinical consistency, stronger representations alone do not prevent the decoder from producing statements that lack direct support in the image. A separate line of work incorporates medical knowledge, disease priors, or graph structures into the generation process to improve clinical correctness (Yang et al., 2022, 2023; Huang et al., 2023; Li et al., 2023a; Huang et al., 2025; Liu et al., 2025). However, their effectiveness depends on the completeness and quality of the injected knowledge that require costly expert annotation. Other approaches (Wang et al., 2024a; Liu et al., 2026) leverage longitudinal patient history to model disease progression, which is clinically valuable but requires prior studies and consistent temporal data. Despite this progress, most existing RRG methods lack explicit mechanisms to trace each generated finding back to inspectable evidence.

Interpretable Radiology Report Generation. A common approach for interpretability in RRG is to incorporate attention maps or saliency-based signals into the generation architecture (Wang et al., 2024b; Chen et al., 2024). Yet recent evidence shows that such signals do not necessarily correspond to the quantities that determine the model prediction, offering an unreliable basis for clinical trust (Salahuddin et al., 2022; Imran and Lee, 2026). One interpretable-by-design alternative routes generation through clinically meaningful intermediate concepts that can be inspected before the final report is produced (Alam et al., 2025; Salmè et al., 2026). Such bottlenecks, however, are semantic rather than spatial: a model may predict the correct disease label without indicating where in the image the evidence lies. Region-guided approaches address this by structuring generation around detected anatomical regions, so that each sentence can be associated with a localized part of the image (Tanida et al., 2023; Bannur et al., 2024). They rely, however, on region-level annotations that are costly and unavailable for most RRG datasets, and they indicate only where a sentence is anchored, not what specific visual pattern within that region supports it. In contrast, prototype-based networks provide intrinsic spatial interpretability without requiring region-level annotations. Rather than explaining a prediction after it has been produced, they derive the prediction from similarities between local image representations and learned visual prototypes. Consequently, the contribution of individual image regions can be inspected as part of the model's forward computation rather than through an auxiliary post-hoc explanation. Prototype-based approaches have been extensively explored for medical image classification (Barnett et al., 2021; Kim et al., 2021; Gallée et al., 2023; Wang et al., 2025a), but their outputs remain fixed clinical labels rather than free-text reports. Within RRG, XPRONET introduces prototypes as shared image-text representations stored in a cross-modal memory (Wang et al., 2022), whereas AP-ISG learns representations as-

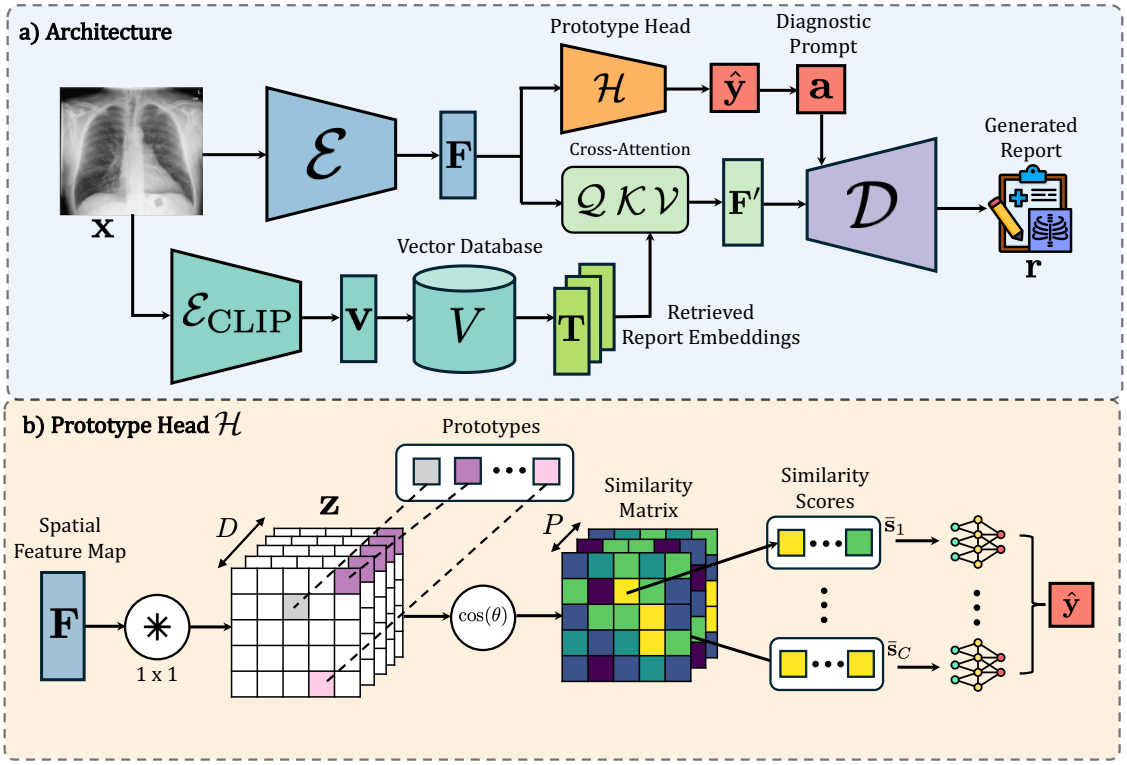


Figure 1: **Overview of the proposed method. (a)** A vision encoder extracts spatial features that are enriched with retrieved report embeddings via cross-attention. A prototype-based head predicts per-condition states, which form a diagnostic prompt for the autoregressive decoder. **(b)** The prototype head computes cosine similarity maps between spatial features and learnable prototypes, providing localised visual evidence for each classification.

sociated with predefined clinical categories (Zhang et al., 2024). In both cases, prototypes or category embeddings support representation learning or generation, but are not matched against local image regions to form the decision for an individual finding. Unlike *ProtoRRG*, neither method matches prototypes to local image regions to directly support the prediction of a specific finding, preventing the resulting prototype matches from being directly localized on the input image.

3 Method

As shown in Fig. 1a, given a chest X-ray image $\mathbf{x} \in \mathbb{R}^{H \times W}$, our goal is to generate a free-text radiology report $\mathbf{r} = (r_1, \ldots, r_T)$ that accurately describes the clinical findings, together with the visual and textual evidence that guided the model. Our method comprises three components. In the first branch, a *prototype-based vision encoder* (Section 3.1) extracts spatial features and classifies each of C thoracic conditions via learnable prototypes that provide inherently interpretable visual evidence. In the second branch, a *retrieval-augmented decoder* (Section 3.2) enriches the visual representation with information from similar reports and generates the output text. The entire model is trained using a *multi-task objective* (Section 3.3) that combines language modelling with classification. This forces the shared vision encoder to support both interpretable diagnosis and fluent report generation simultaneously.

3.1 Prototype-Based Vision Encoder

The vision encoder $\mathcal{E}$ maps the input image $\mathbf{x}$ to a spatial feature map $\mathbf{F} \in \mathbb{R}^{H' \times W' \times D_f}$, which serves a dual role. First, it provides the spatial representations that the decoder $\mathcal{D}$ attends to when generating the report (Section 3.2). Second, it must support the classification of each of the C thoracic conditions, which we address through an inherently interpretable prototype head $\mathcal{H}$ (Fig. 1b). Since the target output is a free-text report, simply knowing which conditions are present is not informative enough: the decoder also needs to know whether each condition is explicitly absent, uncertain, or not mentioned at all. We therefore classify each condition as one of S=4 mutually exclusive states (*positive*, *negative*, *uncertain*, or *blank*), in accordance with established radiology labelling protocols (Smit et al., 2020).

We define $C \times M$ learnable prototypes $\mathbf{P} = \{\mathbf{p}_{c,m} \in \mathbb{R}^D\}_{c=1,m=1}^{C,M}$, associating M prototypes with each of the C conditions. A sequence of 1×1 convolutional layers projects $\mathbf{F}$ to the prototype dimensionality, yielding a spatial feature grid $\mathbf{z} \in \mathbb{R}^{H'\times W'\times D}$. For each spatial location (h, w) of $\mathbf{z}$, we compute the cosine similarity between the corresponding local feature vector $\mathbf{z}_{h,w}$ and each prototype:

$$s_{c,m}(h, w) = \frac{\mathbf{z}_{h,w}^\top \mathbf{p}_{c,m}}{\|\mathbf{z}_{h,w}\| \, \|\mathbf{p}_{c,m}\|} \tag{1}$$

The resulting map $s_{c,m}(\cdot, \cdot)$ scores how closely each image region matches the learned visual pattern for condition c. This makes it possible to identify the areas that contributed the most to a given diagnosis directly from the model's internal representations. Max-pooling over spatial locations then compresses this map into a single score per prototype, $\bar{s}_{c,m} = \max_{h,w} s_{c,m}(h, w)$, selecting the image region that most closely matches each prototype for condition c. Finally, a condition-specific linear layer maps these scores to logits over the S states:

$$\hat{\mathbf{y}}_c = \mathbf{W}_c \, \bar{\mathbf{s}}_c + \mathbf{b}_c \in \mathbb{R}^S \tag{2}$$

where $\mathbf{W}_c \in \mathbb{R}^{S\times M}$ and $\mathbf{b}_c \in \mathbb{R}^S$ are the layer parameters for condition c.

3.2 Retrieval-Augmented Report Generation

Radiology reports tend to follow recurring linguistic patterns, with similar clinical findings often being described using similar phrasing. To exploit this regularity, we retrieve text embeddings from reports associated with visually similar training images and fuse them with the spatial features $\mathbf{F}$ before generation. The retrieved reports can also be examined alongside the generated one, providing additional insight into the similar cases that guided the generation process.

Retrieval and feature enhancement. Before training begins, a frozen vision-language encoder $\mathcal{E}_{\text{CLIP}}$ is used to construct a vector database V: each training image is mapped to a visual embedding $\mathbf{v}_i \in \mathbb{R}^{D_t}$ by the visual branch, and each associated report is mapped to a text embedding $\mathbf{t}_i \in \mathbb{R}^{D_t}$ by the text branch. Given an input image $\mathbf{x}$, the visual branch of $\mathcal{E}_{\text{CLIP}}$ produces an embedding $\mathbf{v}$, and the K training images with the highest cosine similarity to $\mathbf{v}$ in V are retrieved. The text embeddings of their associated reports, $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_K] \in \mathbb{R}^{K\times D_t}$, are then fused with $\mathbf{F}$ via a cross-attention block. Each spatial vector in $\mathbf{F}$ is projected to a D_t-dimensional query, while $\mathbf{T}$ serves as keys and values:

$$\mathbf{F}' = \mathbf{F}\mathbf{W}^{\mathcal{Q}} + \text{CrossAtt}\big(\mathbf{F}\mathbf{W}^{\mathcal{Q}},\ \mathbf{T}\mathbf{W}^{\mathcal{K}},\ \mathbf{T}\mathbf{W}^{\mathcal{V}}\big) \tag{3}$$

This design preserves the spatial structure of the visual representation while allowing each image location to selectively incorporate linguistic context from the retrieved reports. Importantly, the retrieved information is not directly concatenated with the decoder input, but can influence the representation only through cross-attention driven by the spatial image features. In addition, the residual connection explicitly retains the projected visual representation at each spatial position. Together, these architectural choices provide a controlled mechanism for incorporating retrieved information while limiting its direct transfer to the decoder. We quantify the influence of retrieved reports on the generated text in Section F.2.

The resulting representation $\mathbf{F}'$ is used in place of $\mathbf{F}$ as input to the text decoder, conditioning report generation on the visual features augmented with retrieved linguistic context.

Text decoder. The report is generated token by token by an autoregressive language model $\mathcal{D}$, attending to the enhanced features $\mathbf{F}'$ via cross-attention. Following Jin et al. (2024), we condition generation on the classification output by converting each predicted state into a special token that serves as a discrete clinical cue for the decoder. Specifically, the predicted state for each condition c is mapped to a dedicated token $a_c \in \mathcal{A}$, where $\mathcal{A} = \{\texttt{[POS]}, \texttt{[NEG]}, \texttt{[UNC]}, \texttt{[BLK]}\}$ denotes the set of state tokens corresponding to the four mutually exclusive states defined in Section 3.1. The resulting diagnostic prompt $\mathbf{a} = (a_1, \ldots, a_C)$ is prepended to the decoder input, providing an explicit clinical summary that conditions the generation process towards the predicted findings. During training, the diagnostic prompt $\mathbf{a}$ is constructed from the ground-truth state labels via teacher forcing. At inference time, it is derived from the predicted states $\hat{\mathbf{y}}_c$ of the prototype head. Formally:

$$p(r_t \mid r_{<t}, \mathbf{a}, \mathbf{F}') = \mathcal{D}(r_{<t},\ \mathbf{a},\ \mathbf{F}') \tag{4}$$

where each token r_t is generated conditioned on the previously generated tokens, the diagnostic prompt $\mathbf{a}$, and the enhanced visual features $\mathbf{F}'$. The extent

to which the generated report agrees with these diagnostic cues, as well as the effect of replacing predicted states with ground-truth states at inference time, is evaluated in Section F.1.

3.3 Training Objectives and Strategy

The model is trained using a multi-task objective that jointly optimises report generation and per-condition classification.

Training losses. The primary training signal is the standard autoregressive cross-entropy over the generated report:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log\, p(r_t \mid r_{<t}, \mathbf{a}, \mathbf{F}') \tag{5}$$

As radiology labels are severely imbalanced across both pathologies and states, we introduce a Multi-State Focal (MSF) loss function that focuses on each condition and state individually:

$$\mathcal{L}_{\text{MSF}} = -\frac{1}{C}\sum_{c=1}^{C} \alpha_{c,s_c^*} \left(1-p_c(s_c^*)\right)^{\gamma_{c,s_c^*}} \log p_c(s_c^*) \tag{6}$$

where s_c^* is the ground-truth state for condition c, and $p_c(s_c^*)$ denotes the predicted probability of that state. The weighting coefficients $\alpha_{c,s}$ are estimated from empirical class priors as described in Cui et al. (2019), with an additional correction for the *positive* state. The focusing parameters $\gamma_{c,s}$ are set inversely proportional to the same priors, so that training concentrates on the more difficult and rare cases. This formulation extends the asymmetric loss of Ridnik et al. (2021) from multi-label binary classification to the multi-state setting with per-condition parametrisation.

An auxiliary *cluster loss* regularises the prototypes by operating directly on the cosine similarity scores from Eq. (1). As the prototypes are shared across all four states, we introduce state-dependent targets $\tau(s)$ over the $[-1, 1]$ similarity range to differentiate their activations: $\tau(\textit{positive}) = +1$, encouraging high similarity when visual evidence for a condition is present; $\tau(\textit{blank}) = \tau(\textit{negative}) = -1$, as both indicate the absence of supporting evidence, whether unmentioned or explicitly ruled out; and $\tau(\textit{uncertain}) = 0$, capturing equivocal evidence. Let $\sigma_c = \max_m \bar{s}_{c,m}$ denote the maximum prototype similarity for condition c. The cluster loss pulls σ_c towards the target for the ground-truth state:

$$\mathcal{L}_{\text{clst}} = \frac{1}{C}\sum_{c=1}^{C} \big(\tau(s_c^*) - \sigma_c\big)^2 \tag{7}$$

The full training loss combines all terms:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda\big(\mathcal{L}_{\text{MSF}} + \mu\, \mathcal{L}_{\text{clst}}\big) \tag{8}$$

where λ controls the overall weight of the classification branch relative to language modelling, and μ balances the cluster regulariser against the MSF loss within that branch.

Three-phase training. Training proceeds in three phases. First, only the vision backbone $\mathcal{E}$ and the text decoder $\mathcal{D}$ are trained using the language modelling loss $\mathcal{L}_{\text{LM}}$, without the prototype head or classification objectives. This enables the backbone to develop dense spatial features suited to RRG, before the prototype head imposes its own discriminative demands on the same representation. In the second phase, the backbone is frozen and only the projection layers, prototypes, and classification heads are trained with $\mathcal{L}_{\text{MSF}}$ and $\mathcal{L}_{\text{clst}}$. Following Chen et al. (2019), this isolation ensures that the prototypes converge to stable, interpretable patterns in a fixed feature space; training them jointly from initialisation would cause the prototypes and backbone to co-adapt before either has settled, yielding prototypes that lack clear semantic correspondence to image regions. In the third phase, all parameters are updated jointly with the full objective $\mathcal{L}$, using reduced learning rates for the backbone relative to the remaining modules. This staged approach resolves the conflict between the two functions of the shared backbone: the first phase establishes dense spatial features suited to generation; the second phase establishes interpretable prototypes in a stable feature space; and the third phase reconciles both objectives end-to-end.

4 Experimental Setup

4.1 Benchmark Datasets

We conduct experiments on two RRG benchmarks. MIMIC-CXR (Johnson et al., 2019) is a large-scale corpus of radiographs and associated reports collected at the Beth Israel Deaconess Medical Center. We adopt the split of previous works (Chen et al., 2021; Huang et al., 2023; Jin et al., 2024), yielding 270,790 / 2,130 / 3,858 samples for training, validation, and testing. IU X-Ray (Demner-Fushman et al., 2016) provides 3,955 reports covering 7,470

frontal and lateral radiographs from Indiana University. To assess cross-domain generalisability, following prior work (Jin et al., 2024; Zhang et al., 2024), we evaluate the model trained on MIMIC-CXR directly on the IU X-Ray corpus, without fine-tuning.

4.2 Evaluation Metrics

We assess generated reports along two complementary axes. *NLP metrics* include BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), and ROUGE-L (Lin, 2004). *Clinical metrics* are obtained by extracting 14 disease labels from both generated and reference reports with CheXbert (Smit et al., 2020) and computing precision, recall, and F1 over the resulting label sets.

4.3 Prototype Evaluation Protocol

To assess whether the learned prototypes provide spatially faithful evidence, we measure how often their peak activations fall inside expert-curated anatomical regions. The evaluation set is constructed by intersecting the MIMIC-CXR test split with ChestImaGenome (Wu et al., 2021), which provides bounding boxes associating clinical findings with anatomical regions, retaining 2,160 images spanning 5,167 positive (image, class) pairs; full curation details are provided in Appendix A. For each positive pair where the classifier correctly predicts the class, we take the 3 prototypes with the highest activation score $\bar{s}_{c,m}$. The maximally activated location of each prototype's similarity map (Eq. (1)) identifies a single cell on the $H' \times W'$ feature grid, which maps to a rectangular patch P in the input image. Since anatomical bounding boxes B are substantially larger than P, standard intersection-over-union would be dominated by this size disparity rather than by localisation quality. We therefore report an overlap metric,

$$\mathrm{ovl}(P, B) = \frac{|P \cap B|}{|P|}, \tag{9}$$

which reaches 1 when the prototype patch lies entirely inside the ground-truth region. For each prototype we take the maximum overlap over all ground-truth boxes of the class, and the per-image score is the mean of the three overlaps.

4.4 Implementation Details

The vision backbone $\mathcal{E}$ is a ResNet-101 pretrained on ImageNet. We use $M{=}10$ prototypes per condition across $C{=}14$ thoracic conditions. The text decoder is a BERT-base language model initialised from pretrained weights, with cross-attention layers over the encoder output. For retrieval, a frozen BiomedCLIP-CXR (Zambrano Chaves et al., 2025) encoder queries a FAISS index over the training set, retrieving $K{=}3$ neighbours whose text embeddings ($D_t{=}512$) are fused with the backbone features via 8-head cross-attention. We train with AdamW using component-specific learning rates: $5{\times}10^{-5}$ for the backbone and decoder, $1{\times}10^{-4}$ for the projection layers, prototypes, and classification heads, and $5{\times}10^{-5}$ for the generation cross-attention module. The loss weights are $\lambda{=}4.0$ and $\mu{=}0.4$. The MSF parameters $\alpha_{c,s}$ and $\gamma_{c,s}$ are computed once from the training set class priors and kept fixed throughout training; full details are provided in Appendix B. The three training phases run for 1, 5, and 10 epochs respectively, with a batch size of 16.

5 Results and Discussion

5.1 Quantitative Results

To assess our approach, we compare *ProtoRRG* against a broad set of recent state-of-the-art RRG systems (Chen et al., 2020; Li et al., 2023a; Jin et al., 2024; Wang et al., 2025b; Liu et al., 2025), as well as methods that, like ours, provide an interpretability mechanism (Wang et al., 2022; Tanida et al., 2023; Wang et al., 2024b; Chen et al., 2024; Zhang et al., 2024).

On MIMIC-CXR (Table 1), *ProtoRRG* achieves the highest clinical F1 (0.490) and Recall (0.560) across all compared methods. The recall improvement over PromptMRG is particularly notable ($+10\%$ relative, from 0.509 to 0.560), with only a marginal decrease in precision ($0.501 \rightarrow 0.486$). This trade-off is particularly relevant in assistive RRG settings, where generated reports are reviewed by a radiologist: higher recall can reduce the risk of omitting correct findings, although it may come at the cost of an increased number of false positives. In our formulation, this balance is explicitly controlled by the positive-state scaling factor in the training objective, whose effect on precision and recall is analysed in Table 7. Among interpretable models, the margin widens further: clinical F1 improves by 6.1% over AP-ISG (0.462) and by 9.6% over RGRG (0.447). These results indicate that visual prototypes not only provide spatially grounded evidence for predicted findings, but also improve the coverage of clinically rele-

Table 1: Comparison on the MIMIC-CXR test set. † denotes methods evaluated on a different data split. $*$ denotes results recomputed from the official checkpoint using the same preprocessing and evaluation pipeline adopted for *ProtoRRG*, as the published values were obtained under a different setting. Results for *ProtoRRG* are averaged over three training runs with different random seeds; a statistical analysis of the results is reported in Appendix C. Best results are shown in **bold** and second-best results are underlined.

	Clinical Metrics			NLP Metrics			
Model	**P**	**R**	**F1**	**B-1**	**B-4**	**MTR**	**RG-L**
Non-interpretable methods							
R2Gen (Chen et al., 2020)	0.333	0.273	0.276	0.353	0.103	0.142	0.277
M^2TR (Cornia et al., 2020)	0.240	0.428	0.308	0.378	0.107	0.145	0.272
CliBert (Yan and Pei, 2022)	0.397	0.435	0.415	0.383	0.106	0.144	0.275
CVT2Dis. (Nicolson et al., 2023)	0.356	0.412	0.384	0.392	0.124	0.153	0.285
M^2KT (Yang et al., 2023)	0.420	0.339	0.352	0.386	0.111	–	0.274
ME (Wang et al., 2023)	0.364	0.309	0.311	0.386	0.124	0.152	0.291
KiUT (Huang et al., 2023)	0.371	0.318	0.321	0.393	0.113	0.160	0.285
DCL (Li et al., 2023a)	0.471	0.352	0.373	–	0.109	0.150	0.284
HERGen (Wang et al., 2024a)	0.415	0.301	0.317	0.395	0.122	0.156	0.285
PromptMRG (Jin et al., 2024)	**0.501**	0.509	0.476	0.398	0.112	0.157	0.268
R2Gen-Mamba (Sun et al., 2025)	0.483	0.325	0.389	0.352	0.110	0.141	0.284
MambaXray-VL-L* (Wang et al., 2025b)	0.446	0.342	0.364	0.403	0.125	0.160	0.274
SLGT (Liu et al., 2025)	0.461	0.356	0.375	0.401	0.122	0.168	**0.295**
DACG (Lang et al., 2025)	0.422	0.405	0.389	0.398	0.117	0.162	0.290
Interpretable methods							
XPRONET (Wang et al., 2022)	0.463	0.285	0.353	0.344	0.105	0.138	0.279
CAMANet (Wang et al., 2024b)	0.483	0.323	0.387	0.374	0.112	0.145	0.279
AdaMatch-Cyclic (Chen et al., 2024)	–	–	–	0.379	0.106	0.162	0.286
RGRG† (Tanida et al., 2023)	0.461	0.475	0.447	0.373	0.126	0.168	0.264
AP-ISG† (Zhang et al., 2024)	0.486	0.493	0.462	0.391	**0.129**	**0.175**	0.282
ProtoRRG (Ours)	0.486	**0.560**	**0.490**	**0.404**	0.110	0.161	0.272

vant conditions in the generated reports. On NLP metrics, *ProtoRRG* achieves the highest BLEU-1 (0.404), marginally above the strongest recent baselines, while BLEU-4, METEOR, and ROUGE-L remain competitive but are not uniformly best. We interpret this pattern as stronger recovery of clinically salient lexical units, rather than long-range n-gram overlap. This behaviour is consistent with our objective design, which prioritises clinical faithfulness and finding coverage through prototype-guided state prompting. In clinical report generation, this trade-off is desirable when it improves diagnostic content.

To assess cross-domain robustness, we apply the MIMIC-CXR-trained model directly to IU X-Ray without fine-tuning (Table 2). *ProtoRRG* achieves the best performance across all clinical metrics, with P 0.222, R 0.236, and F1 0.220. Its clinical F1 exceeds that of the strongest interpretable baseline, AP-ISG (0.189), by 16%, and also surpasses the best non-interpretable baseline, PromptMRG (0.211). A similar trend is observed for the NLP metrics: *ProtoRRG* obtains the highest BLEU-1 (0.418), BLEU-4 (0.102), and ROUGE-L (0.308), while its METEOR score (0.159) is marginally below the best reported result. These results indicate that the model transfers both finding-level content and report phrasing to a new domain, suggesting that it captures patterns beyond MIMIC-CXR-specific regularities.

In summary, *ProtoRRG* improves clinically meaningful content in an interpretable manner, while maintaining competitive report-level language quality.

5.2 Ablation Study

To isolate the contribution of each component, we start from an encoder-decoder baseline trained with only $\mathcal{L}_{\text{LM}}$ and progressively add one module per row (Table 3).

Adding the prototype head with standard cross-entropy produces the single largest gain: clinical F1 rises from 0.375 to 0.452 (+20.5% relative) and Recall from 0.385 to 0.556 (+44.4%). This confirms that the prototype head is not merely an interpretability mechanism: the diagnostic prompt $\mathbf{a}$ actively prevents the decoder from omitting find-

Table 2: Cross-domain evaluation on the IU X-Ray dataset. All models are trained on MIMIC-CXR and evaluated without fine-tuning. Results for *ProtoRRG* are averaged over the same three training runs. Best results are shown in **bold** and second-best results are underlined.

	Clinical Metrics			NLP Metrics			
Model	**P**	**R**	**F1**	**B-1**	**B-4**	**MTR**	**RG-L**
Non-interpretable methods							
R2Gen (Chen et al., 2020)	0.141	0.136	0.136	0.325	0.059	0.131	0.253
CVT2Dis. (Nicolson et al., 2023)	0.174	0.172	0.168	0.383	0.082	0.147	0.277
M^2KT (Yang et al., 2023)	0.153	0.145	0.145	0.371	0.078	0.153	0.261
DCL (Li et al., 2023a)	0.168	0.167	0.162	0.354	0.074	0.152	0.267
PromptMRG (Jin et al., 2024)	0.213	0.229	0.211	0.401	0.098	**0.160**	0.281
Interpretable methods							
RGRG (Tanida et al., 2023)	0.183	0.187	0.180	0.266	0.063	0.146	0.180
AP-ISG (Zhang et al., 2024)	0.192	0.194	0.189	–	0.076	0.148	–
ProtoRRG (Ours)	**0.222**	**0.236**	**0.220**	**0.418**	**0.102**	0.159	**0.308**

Table 3: Progressive ablation on the MIMIC-CXR test set. Starting from an encoder-decoder baseline, each row adds one component: the prototype head with cross-entropy (Proto.), the three-phase training strategy (3-phase), the MSF loss replacing cross-entropy (MSF), and Retrieval-Augmented Generation (RAG). All configurations correspond to a single training run.

Model	**Clinical Metrics**			**NLP Metrics**		
	P	R	F1	B-1	B-4	RG-L
Baseline	0.421	0.385	0.375	0.363	0.103	0.270
+ Proto.	0.433	0.556	0.452	0.376	0.093	0.255
+ 3-phase	0.475	0.516	0.464	0.396	0.106	0.265
+ MSF	0.474	0.571	0.488	0.398	0.105	0.266
+ RAG (*full*)	0.487	0.556	0.490	0.404	0.111	0.273

Table 4: Prototype localisation evaluation on MIMIC-CXR. N: number of positive instances; **Rec.**: classification recall; **Avg Ovl**: mean overlap between the top-3 prototype patches and the ground-truth bounding boxes; $|B|$: mean bounding-box area as a fraction of the image.

Class	N	**Rec.**	**Avg Ovl**	$\lvert B\rvert$ (%)
Atelectasis	689	0.759	0.696	21.4
Cardiomegaly	843	0.823	0.921	20.2
Consolidation	118	0.475	0.768	19.5
Edema	435	0.834	0.730	17.5
Enl. Cardiomed.	123	0.244	0.912	18.7
Fracture	28	0.000	–	18.1
Lung Lesion	134	0.060	0.793	20.8
Lung Opacity	949	0.758	0.812	16.6
Pleural Effusion	814	0.778	0.724	16.6
Pleural Other	71	0.338	0.679	21.2
Pneumonia	122	0.549	0.789	20.7
Pneumothorax	55	0.582	0.701	22.7
Support Devices	786	0.977	0.918	22.4
Overall	5167	0.758	0.813	18.7

ings. Introducing the three-phase training strategy further improves F1 from 0.452 to 0.464, with a notable effect on NLP metrics, indicating that staged training helps the shared backbone better balance discriminative and generative demands. Replacing cross-entropy with the MSF loss yields a further improvement in clinical F1 to 0.488, confirming the benefit of dedicated handling of class-state imbalance. In the final configuration, adding retrieval results yields gains on both axes (F1: 0.490; B-1: 0.404; B-4: 0.111), with larger gains on lexical metrics than clinical ones. This suggests that there is a complementary division of labour: prototype-guided prompting primarily improves diagnostic coverage, while retrieval primarily improves lexical formulation and phrasing. In summary, each component targets a distinct aspect of report quality, and their combination yields the best results on both clinical and lexical metrics.

5.3 Prototype Evaluation

Table 4 shows that *ProtoRRG* attains an overall overlap of 0.813: on average, more than 81% of each prototype patch falls inside the anatomical region associated with the predicted finding. The overlap is not a proxy for bounding-box size; Atelectasis and Support Devices have comparable areas ($|B| = 21.4\%$ and 22.4%) yet differ substantially in overlap (0.696 vs. 0.918). Conditions with stable anatomical correlates (Cardiomegaly, Enlarged Cardiomediastinum, Support Devices) exhibit the strongest grounding (> 0.91), while spatially diffuse conditions (Atelectasis, Pleural Effusion, Pleural Other) yield more moderate overlaps (0.68–0.73). Fracture (N=28) is never predicted positive on this subset and Lung Lesion

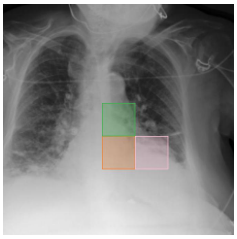
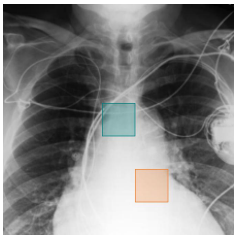
Image & Prototypes	Retrieved Report (top-1)	Generated Report	Ground truth
	*Portable AP upright chest radiograph was obtained. Compared with the previous examination a similar degree of perihilar opacity with slight decrease in interstitial thickening is consistent with moderate pulmonary edema. The density of the left lower lobe/retrocardiac opacity is such that an infectious consolidation may be obscured in this area. Bilateral pleural effusions are unchanged. The heart remains mild to moderately enlarged with calcified aortic knob.*	The **heart size is mildly enlarged**. The aorta is tortuous and diffusely calcified. There is perihilar haziness and vascular indistinctness compatible with mild **pulmonary edema**. Small bilateral **pleural effusions** left greater than right are noted. No pneumothorax is identified. Percutaneous gastrostomy catheter is seen within the left upper quadrant of the abdomen. There are no acute osseous abnormalities.	The **heart size is mildly enlarged**. The aortic knob is calcified. There is mild **pulmonary edema** noted with perihilar haziness and vascular indistinctness. Small bilateral **pleural effusions** are present left greater than right with bibasilar airspace opacities. No pneumothorax is identified. There are no acute osseous abnormalities.
	*There is moderate cardiomegaly. Left chest wall dual lead pacing device seen with tips projecting over the right ventricle and right atrium. The lungs are clear where not obscured by overlying left chest wall pacing device, without focal consolidation, large effusion or edema. No visualized acute osseous abnormalities.*	AP portable upright view of the chest. Left chest wall pacer **device** is again seen with leads extending to the region the right atrium and right ventricle. The **heart remains mildly enlarged**. There is no focal consolidation effusion or pneumothorax. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen. Aortic calcification again noted.	AP upright portable chest radiograph is obtained. A left chest wall pacer **device** is again seen with lead tips extending into the right atrium and ventricle. Abandoned pacing leads are also seen in the right chest wall extending into the right heart not significantly changed. The **heart is mildly enlarged**. The lungs appear clear without definite signs of pneumonia or CHF. No large effusion or pneumothorax is seen. The overall cardiomediastinal silhouette is stable. Bony structures are intact.

Figure 2: Qualitative examples on two MIMIC-CXR test cases. For each case, the input radiograph is shown with the activated prototypes, followed by the top-1 retrieved report, the generated report, and the ground truth. Colours mark findings supported by the ground truth, and show how each of them is traceable to both a localised prototype activation and a case-based linguistic precedent.

(Rec. $= 0.060$) yields too few correct detections for a stable localisation estimate; both findings are known to challenge deep-learning classifiers owing to extreme class imbalance and subtle radiographic presentation (Lin et al., 2023; Guo et al., 2024). With these two exceptions, the results support the claim that the learned prototypes surface spatially faithful visual evidence for the findings they drive.

6 Conclusion

We introduced *ProtoRRG*, a prototype-based framework for RRG that grounds each finding in localised prototype activations and retrieved prior reports. A shared spatial representation reconciles prototype-based classification with autoregressive decoding, trained end-to-end through a three-phase strategy. Experiments on MIMIC-CXR and IU X-Ray show that *ProtoRRG* achieves state-of-the-art clinical accuracy while outperforming every prior interpretable RRG system, and an anatomically grounded evaluation confirms that its prototypes consistently localise to the correct regions. These results suggest that interpretability and diagnostic accuracy need not be traded off against each other in RRG.

Limitations

Despite the strong results across both clinical and interpretability evaluations, several limitations warrant discussion. First, each prototype identifies a single fixed-size region as evidence for a given finding, which inherently limits spatial grounding for conditions that are diffuse or smaller than the feature-grid resolution. Second, the prototypes are learned entirely from image-level labels without any radiologist feedback; incorporating even a small set of expert spatial annotations or physician-in-the-loop training could better align the learned evidence patterns with clinical reasoning. Our evaluation, moreover, relies on expert-curated annotations rather than on direct radiologist assessment of the generated reports, which would provide complementary evidence of clinical utility. Finally, all experiments are restricted to chest radiographs, and whether the learned prototypes generalise to other imaging modalities or anatomical regions remains untested. Addressing these limitations constitutes a natural direction for future work.

Acknowledgments

Marco Salmè is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXIX cycle, course on Health and Life Sciences, organized by Università Campus Bio-Medico di Roma. This work was partially funded by: i) Università Campus Bio-Medico di Roma under the program "University Strategic Projects" within the project "AI-powered Digital Twin for next-generation lung cancEr cAre (IDEA)"; ii) Cancerforskningsfonden Norrland LP 25-2383; iii)

Kempe Foundation project JCSMK24-0094; iv) Fondo di Beneficenza Intesa San Paolo - project ID: B/2025/0360. Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Alvis @ C3SE, partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973.

References

Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. 2025. Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag. In *European Conference on Information Retrieval*, pages 201–209. Springer.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. 2021. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070.

Simon Baur, Wojciech Samek, and Jackie Ma. 2025. Benchmarking uncertainty and its disentanglement in multi-label chest x-ray classification. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 193–203. Springer.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and 1 others. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, volume 32.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9494–9509.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449. Association for Computational Linguistics.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10575–10584. IEEE.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9260–9269. IEEE.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest X-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health (ML4H), Proceedings of Machine Learning Research*, pages 209–219. PMLR.

Luisa Gallée, Meinrad Beer, and Michael Götz. 2023. Interpretable medical image classification using prototype learning and privileged information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 435–445. Springer.

Lili Guo, Changsheng Zhou, Jingxu Xu, Chencui Huang, Yizhou Yu, and Guangming Lu. 2024. Deep learning for chest x-ray diagnosis: competition between radiologists with or without artificial intelligence assistance. *Journal of Imaging Informatics in Medicine*, 37(3):922–934.

Lili Huang, Yiming Cao, Pengcheng Jia, Chenglong Li, Jin Tang, and Chuanfu Li. 2025. Knowledge-guided cross-modal alignment and progressive fusion for chest x-ray report generation. *IEEE Transactions on Multimedia*, 27:557–567.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19809–19818. IEEE.

Muhammad Imran and Yugyung Lee. 2026. Multimodal vision–language models in medical imaging: A survey of retrieval, interpretability, and trust. *IEEE Access*.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2607–2615.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. 2021. Xprotonet: diagnosis in chest radiography with global and local explanations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15714–15723. IEEE.

Wangyu Lang, Zhi Liu, and Yijia Zhang. 2025. Dacg: Dual attention and context guidance model for radiology report generation. *Medical Image Analysis*, 99:103377.

Dongfang Li, Hua Huo, Shupei Jiao, Xiaowei Sun, and Shuya Chen. 2024. Automated thorax disease diagnosis using multi-branch residual attention network. *Scientific Reports*, 14(1):11865.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023a. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343. IEEE.

Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023b. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2851–2862. IEEE.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Mingquan Lin, Bojian Hou, Swati Mishra, Tianyuan Yao, Yuankai Huo, Qian Yang, Fei Wang, George Shih, and Yifan Peng. 2023. Enhancing thoracic disease detection using chest x-rays from pubmed central open access. *Computers in biology and medicine*, 159:106962.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. 2026. Priorrg: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7206–7214.

Xinyao Liu, Junchang Xin, Qi Shen, Zhihong Huang, and Zhiqiong Wang. 2025. Soft label-guided transformer for radiology report generation. *Journal of Biomedical Informatics*, page 104924.

Xin Mei, Libin Yang, Dehong Gao, Xiaoyan Cai, Junwei Han, and Tianming Liu. 2025. Adaptive medical topic learning for enhanced fine-grained cross-modal alignment in medical report generation. *IEEE Transactions on Multimedia*, 27:5050–5061.

Meike Nauta, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2744–2753. IEEE.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91. IEEE.

Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. 2022. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111.

Marco Salmè, Federico Siciliano, Fabrizio Silvestri, Paolo Soda, Rosa Sicilia, and Valerio Guarrasi. 2026. Concept-enhanced multimodal rag: Towards interpretable and accurate radiology report generation. *Information Systems Frontiers*, pages 1–26.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1500–1519.

Steven Song, Jiaying Fang, Stephanie Reis, and Ronnachai Jaroensri. 2024. LaB-RAG: Label boosted retrieval augmented generation for radiology report generation. *arXiv preprint arXiv:2411.16523*.

Yongheng Sun, Yueh Z Lee, Genevieve A Woodard, Hongtu Zhu, Chunfeng Lian, and Mingxia Liu. 2025. R2gen-mamba: A selective state space model for radiology report generation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7433–7442. IEEE.

Chong Wang, Fengbei Liu, Yuanhong Chen, Helen Frazer, and Gustavo Carneiro. 2025a. Cross-and intra-image prototypical learning for multi-label disease diagnosis and interpretation. *IEEE Transactions on Medical Imaging*, 44(6):2568–2580.

Fuying Wang, Shenghui Du, and Lequan Yu. 2024a. Hergen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, pages 183–200. Springer.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022. Cross-modal prototype driven network for radiology report generation. In *European conference on computer vision*, pages 563–579. Springer.

Jun Wang, Abhir Bhalerao, Terry Yin, Simon See, and Yulan He. 2024b. Camanet: class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2199–2210.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. 2025b. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5123–5133. IEEE.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11558–11567. IEEE.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, and 1 others. 2021. Chest imagenome dataset. *Physio Net*.

Bin Yan and Mingtao Pei. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2982–2990.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, and 1 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

Ke Zhang, Yan Yang, Jun Yu, Jianping Fan, Hanliang Jiang, Qingming Huang, and Weidong Han. 2024. Attribute prototype-guided iterative scene graph for explainable radiology report generation. *IEEE transactions on medical imaging*, 43(12):4470–4482.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2025. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640.

A Prototype Evaluation Details

A.1 ChestImaGenome Evaluation Set

The localisation evaluation described in Section 4.3 requires images with both CheXpert-style disease labels and spatially grounded anatomical annotations. We obtain these by intersecting the MIMIC-CXR test split with ChestImaGenome (Wu et al., 2021), which provides bounding boxes associating clinical findings with anatomical regions.

ChestImaGenome defines 162 attribute strings at a finer granularity than the 14 CheXpert conditions used by our classifier. We map these attributes to the 13 localisable CheXpert classes (*No Finding* is excluded as it lacks a fixed anatomical region) through a rule-based assignment that groups ImaGenome attributes by their associated

Table 5: Complete ImaGenome-to-CheXpert attribute mapping. Attributes marked with † are assigned to more than one class.

CheXpert Class	ImaGenome Attribute Strings
Enlarged Cardiomediastinum	mediastinal widening/enlargement, widened mediastinum, enlarged cardiac silhouette
Cardiomegaly	cardiomegaly, cardiac enlargement
Lung Opacity	lung opacity, opacity, airspace opacity/disease, hazy opacity†, infiltrate/infiltration, breast/nipple shadows
Lung Lesion	lung lesion/mass, nodule/mass, pulmonary nodule/mass, calcified nodule/granuloma, granuloma, lung cancer
Edema	pulmonary edema, edema, vascular/pulmonary vascular congestion, interstitial/alveolar edema, fluid overload (heart failure), vascular redistribution, cephalization, interstitial thickening/prominence, vascular fullness, perihilar haze, hazy opacity†
Consolidation	consolidation, consolidative opacity, lobar consolidation
Pneumonia	pneumonia
Atelectasis	atelectasis, atelectatic, collapse, lobar/lung/lobar-segmental collapse, linear/patchy/subsegmental/bibasilar/basilar/band-like/platelike/discoid atelectasis, volume loss
Pneumothorax	pneumothorax, ptx, hydropneumothorax, tension pneumothorax
Pleural Effusion	pleural effusion, effusion, pleural fluid, hydrothorax, layering/loculated effusion, costophrenic angle blunting
Pleural Other	pleural thickening/scarring/plaque/calcification/reaction, pleural-parenchymal scarring, apical pleural thickening/scarring
Fracture	fracture, rib/clavicular/vertebral/sternal/scapular/spinal fracture, compression fracture, healing/healed/old/acute fracture
Support Devices	*Tubes:* ET/tracheostomy/NG/OG/feeding/enteric/drainage tube, dobhoff, drain. *Lines:* central venous catheter/line, PICC, Swan–Ganz, pulmonary artery/dialysis catheter, port-a-cath, arterial/venous line. *Cardiac:* pacemaker, pacer, AICD, ICD, defibrillator, pacemaker lead(s), cardiac leads/wires. *Surgical:* IABP, LVAD, ECMO, clips, staples, prosthetic valve, CABG, stent, hardware, sternotomy wires.

anatomical region and clinical semantics. The initial mapping was generated by Claude Opus 4.6 and subsequently reviewed and corrected by a radiologist; Table 5 reports the final assignment. To verify the quality of this mapping, we compare the mapped labels against the image-level CheXpert labels extracted by CheXbert (Smit et al., 2020), obtaining an F1 score of 0.935.

We apply two filters to ensure evaluation reliability. First, we discard images that lack any spatial annotation in ChestImaGenome, as these provide no ground-truth regions for localisation. Second, we require that every positive CheXpert label for a given image is backed by at least one affirmed ImaGenome attribute with an associated bounding box. After filtering, we retain 2,160 images spanning 5,167 positive (image, class) pairs across the 13 localisable conditions.

A.2 Additional Localisation Evaluation on MS-CXR

To complement the ChestImaGenome evaluation, we conduct an additional localisation assessment on MS-CXR (Boecking et al., 2022), an independent dataset of 1,047 chest radiographs providing 1,162 bounding-box annotations across 8 cardiopulmonary findings. After intersecting MS-CXR with the MIMIC-CXR test split, we retain 596 positive (image, class) pairs and apply the same protocol described in Section 4.3.

Table 6 reports the results. Without any spatial supervision, the prototypes achieve a mean overlap of 0.473 across the eight findings. Cardiomegaly, which is large and anatomically consistent, reaches near-perfect localisation (0.906). Parenchymal conditions such as Atelectasis, Consolidation, and Edema show overlaps in the range 0.21–0.30, with prototypes activating on diagnostically relevant patterns near the annotated regions. For Consolidation and Pneumonia, classification recall is low (0.370 and 0.214); yet when correctly identified, the associated prototypes localise with reasonable spatial fidelity (0.297 and 0.496), suggesting that the primary bottleneck is detection rather than spatial precision. Pneumothorax and Pleural Effusion yield the lowest overlaps (0.081 and 0.078), consistent with their small annotation

Table 6: Prototype localisation on MS-CXR. N: positive instances; **Rec.**: classification recall; **Avg Ovl**: mean overlap (Eq. (9)); $|\mathcal{B}|$: mean bounding-box area as a fraction of the image.

Finding	N	Rec.	Avg Ovl	$\lvert\mathcal{B}\rvert$ (%)
Atelectasis	41	0.902	0.273	14.8
Cardiomegaly	189	0.952	0.906	26.2
Consolidation	73	0.370	0.297	17.8
Edema	32	0.969	0.210	30.9
Lung Opacity	42	0.929	0.173	14.7
Pleural Effusion	51	0.922	0.078	10.6
Pneumonia	56	0.214	0.496	16.6
Pneumothorax	112	0.634	0.081	7.2
Overall	596	0.745	0.473	18.0

areas ($|\mathcal{B}|$=7.2% and 10.6%): the coarse spatial resolution of the feature map limits precise alignment with compact targets. In summary, prototype localisation is reliable across the majority of conditions, with accuracy decreasing mainly for findings that are spatially compact or radiographically subtle.

B Hyperparameters and Training Details

This appendix complements the implementation details in Section 4.4 with additional training configuration and loss parameterisation.

Training configuration. All phases use AdamW with weight decay 1×10^{-3}, a linear warmup of 2,000 steps, cosine annealing to a minimum learning rate of 5×10^{-6}, and images resized to 224×224 pixels. In Phase 1, only the backbone $\mathcal{E}$ and decoder $\mathcal{D}$ are trained with $\mathcal{L}_{\text{LM}}$; the prototype head is inactive. In Phase 2, $\mathcal{E}$ is frozen and only the projection layers, prototypes $\mathbf{P}$, and classification heads are updated with $\mathcal{L}_{\text{MSF}} + \mu\,\mathcal{L}_{\text{clst}}$. In Phase 3, all parameters are updated jointly with the full objective $\mathcal{L}$ (Eq. (8)), using the component-specific learning rates reported in Section 4.4. The checkpoint with the highest CheXbert F1 on the validation set is retained for testing. On a single NVIDIA A40 GPU, the three phases take approximately 2 hours, 45 minutes, and 18 hours respectively, for a total training time of roughly 21 hours.

MSF loss parameters. The per-condition weights $\alpha_{c,s}$ are obtained via effective-number balancing with β=0.9999, which Cui et al. (2019) find optimal for coarse-grained tasks with strong class imbalance. Weights are normalised so that $\sum_s \alpha_{c,s} = S$ for every condition c, with an additional multiplicative factor of 2.0 on the *positive* state to counteract the dominance of *blank*. The focusing exponents $\gamma_{c,s}$ are set inversely proportional to within-condition class frequency and scaled to $[0, 2]$, following the search range of Cui et al. (2019) and Lin et al. (2017). The *positive* state uses a fixed γ=0.5, consistent with the default recommended by Cui et al. (2019) for class-balanced focal loss and with the low γ^+ regime advocated by Ridnik et al. (2021) to preserve gradient signal on rare classes. Both $\alpha_{c,s}$ and $\gamma_{c,s}$ are computed once from training-set label counts and kept fixed throughout all phases. The classification-branch weight λ=4.0 follows Jin et al. (2024). The cluster loss weight μ=0.4 is selected via the sensitivity analysis in Table 7b, which also reports the effect of the positive-state scaling factor α^+ (Table 7a). As shown in Table 7a, increasing α^+ from 1 to 5 improves recall from 0.481 to 0.719, while decreasing precision from 0.497 to 0.361. This trend indicates that α^+ controls the precision-recall trade-off of the prototype head by increasing the emphasis on positive findings. The selected value, α^+=2, achieves the highest F1 score, providing the best balance between finding coverage and false-positive rate.

Table 7: Sensitivity analysis on the MIMIC-CXR test set (classification metrics from the prototype head). **(a)** Positive-state scaling factor α^+ with μ=0.4. **(b)** Cluster loss weight μ with α^+=2.

(a) Scaling factor α^+				**(b) Cluster weight μ**			
α^+	P	R	F1	μ	P	R	F1
1	0.497	0.481	0.458	0	0.451	0.628	0.494
2	**0.495**	**0.606**	**0.513**	0.1	0.466	0.608	0.495
3	0.429	0.670	0.493	0.2	0.479	0.609	0.504
4	0.388	0.719	0.479	0.3	0.485	0.610	0.509
5	0.361	0.719	0.465	**0.4**	**0.495**	**0.606**	**0.513**
				0.5	0.492	0.601	0.510

C Statistical Analysis of Results

To assess the reliability of the results reported in Table 1, we train the full model three times with different random seeds, keeping all hyperparameters and the training schedule unchanged. Table 8 reports, for each metric, the mean and standard deviation across the three runs, together with a 95% bootstrap confidence interval obtained by resampling the MIMIC-CXR test set 1,000 times with replacement. The two quantities capture different sources of variation: the standard deviation reflects the sensitivity of the model to random initialisation and data ordering, whereas the bootstrap interval reflects the sensitivity of each metric to the composition of the test set for a fixed model, and is computed on a single run.

Across all metrics the standard deviation does not exceed 0.006, indicating that the model is largely insensitive to random initialisation. The bootstrap intervals are similarly narrow, spanning at most 0.023 on clinical metrics and 0.006 on lexical ones, indicating that the reported values are stable with respect to the composition of the test set.

Table 8: Statistical analysis of *ProtoRRG* results on the MIMIC-CXR test set. Mean and standard deviation are computed over three training runs with different random seeds; the 95% bootstrap confidence interval is obtained from 1,000 resamples of the test set for a single run.

Metric	Mean ± Std	95% CI
BLEU-1	0.404 ± 0.001	$[0.400, 0.408]$
BLEU-4	0.110 ± 0.001	$[0.108, 0.114]$
METEOR	0.161 ± 0.001	$[0.158, 0.164]$
ROUGE-L	0.272 ± 0.001	$[0.271, 0.276]$
Precision	0.486 ± 0.002	$[0.477, 0.498]$
Recall	0.560 ± 0.006	$[0.544, 0.567]$
F1	0.490 ± 0.001	$[0.480, 0.500]$

D Effect of the Vision Backbone

To assess the impact of the vision backbone on *ProtoRRG*, we replace the ResNet-101 encoder with BiomedCLIP ViT-B/16 (Zhang et al., 2025), keeping the prototype head, the retrieval module, and the training strategy unchanged.

Table 9: Effect of the vision backbone on the MIMIC-CXR test set. All other components and training settings are held fixed. Both configurations correspond to a single training run.

Metric	ResNet-101	BiomedCLIP ViT-B/16
Precision	0.487	0.473
Recall	0.556	0.557
F1	0.490	0.482
BLEU-1	0.404	0.400
BLEU-4	0.111	0.108
METEOR	0.161	0.158
ROUGE-L	0.273	0.271

As reported in Table 9, the two backbones achieve comparable recall, while ResNet-101 attains higher precision and F1. This outcome is consistent with the requirements of our architecture, which relies on fine-grained multi-label state classification rather than on coarse image-text alignment. The contrastive pretraining objective of CLIP-style encoders favours global representations, which are less directly aligned with the condition-state discrimination performed by the prototype head. Residual convolutional backbones have been shown to remain competitive with vision foundation models in this setting (Baur et al., 2025; Li et al., 2024). Since the predicted states condition the decoder through the diagnostic prompt, classification performance is a primary criterion when selecting the encoder, which motivates our choice of ResNet-101.

E Additional Qualitative Examples and Failure Modes

To complement the examples in Fig. 2, Fig. 3 presents three additional MIMIC-CXR test cases that illustrate both strengths and failure modes of *ProtoRRG*.

The first case shows a focal pneumonia in the right middle lobe. The generated report correctly identifies both findings and closely matches the ground truth. Notably, the Lung Opacity and Pneumonia prototypes activate on the same spatial region, consistent with the clinical co-occurrence of these conditions at the same anatomical site, confirming that the model grounds semantically related findings in a shared location without conflating their diagnostic labels.

The second case reveals a dissociation between the prototype head and the decoder. The classifier correctly predicts Cardiomegaly and Enlarged Cardiomediastinum with well-localised activations, but does *not* predict Edema as positive. Despite the absence of prototype evidence, the decoder generates mild interstitial pulmonary edema, a finding consistent with the ground truth, likely recovered from the dense visual features alone. This improves report completeness but introduces an interpretability gap: the edema mention lacks the localised visual justification that supports the other findings.

The third case involves four concurrent findings and produces a report closely aligned with the ground truth. Two failure modes are however apparent. First, the retrieved neighbour mentions severe degenerative changes of the right glenohumeral joint; the decoder propagates a variant of this observation (degenerative changes [...] glenohumeral and acromioclavicular joints) that is absent from the ground truth, constituting a retrieval-induced hallucination. Second, the Atelectasis prototype activates over the cardiac region rather than at the lung bases where the ground truth places the finding, reflecting the spatial precision limitation of the fixed feature-grid resolution discussed in Section 6.

F Quantitative Failure Mode Analysis

Appendix E illustrates two failure modes qualitatively: the decoder producing findings absent from the prototype-head prediction (Case 2), and the potential propagation of content from retrieved reports into the generated text (Case 3). This ap-

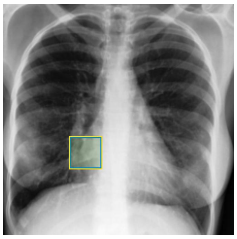
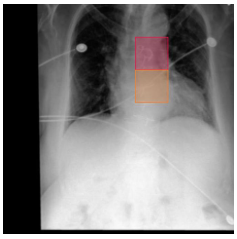
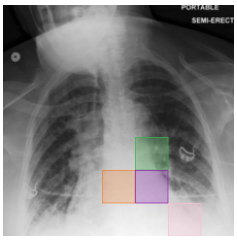
Image & Prototypes	Retrieved Report (top-1)	Generated Report	Ground truth
	*The heart size is normal. The hilar and mediastinal contours are normal. There is a focal consolidation in the right middle lobe concerning for pneumonia. There is no pleural effusion. There is no pneumothorax. The visualized osseous structures are unremarkable.*	Cardiac mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Patchy opacity is noted within the right middle lobe concerning for pneumonia. Left lung is clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities. Mild to moderate multilevel degenerative changes are noted in the thoracic spine.	The heart is normal in size. The mediastinal and hilar contours appear within normal limits. Pleural effusions have resolved. There is a patchy new opacity in the lateral segment of the right middle lobe worrisome for pneumonia. However elsewhere the lungs appear clear. The osseous structures are unremarkable.
	*The lower trachea is deviated to the right secondary to a possibly enlarged, calcified aortic arch. The lungs are grossly clear of consolidation or other evidence of pneumonia. A 9 mm nodule to the left of the cardiac apex might be calcified. There is no pneumothorax or pleural effusion. The patient is status post median sternotomy and CABG. Cardiac silhouette is moderately enlarged, with the particularly dilated right ventricle.*	The patient is status post median sternotomy and CABG. The heart is mildly enlarged. The aorta is tortuous and diffusely calcified. There is mild interstitial pulmonary edema. No focal consolidation pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities. Mild loss of height of a mid thoracic vertebral body is unchanged.	The heart remains enlarged. The aorta is markedly tortuous. Increased interstitial markings are seen throughout the lungs similar to prior and most compatible with edema. No pneumothorax or consolidation or pleural effusion. There is diffuse demineralization. Incidental note is made of a right cervical rib. Sternotomy sutures project over the mediastinum. EKG leads overlie the chest wall.
	*The heart size is enlarged. The mediastinal and hilar contours are stable. There has been an increase in pulmonary edema. Small bilateral pleural effusions with basilar atelectasis are also present. There is no pneumothorax. Severe degenerative changes of the right glenohumeral joint are visualized.*	Lung volumes are low. Heart size is mildly enlarged. The aorta is tortuous and demonstrates atherosclerotic calcifications. There is mild pulmonary edema with small bilateral pleural effusions. Bibasilar patchy opacities likely reflect areas of atelectasis. No pneumothorax is identified. There are no acute osseous abnormalities. Degenerative changes are noted involving both glenohumeral and acromioclavicular joints.	The heart size appears moderately enlarged. The mediastinum demonstrates tortuosity of the thoracic aorta. There is perihilar haziness with vascular indistinctness compatible with mild pulmonary edema. Hazy opacities in both lung bases likely reflect small layering bilateral pleural effusions with associated bibasilar atelectasis. No large pneumothorax is identified. There are no acute osseous abnormalities.

Figure 3: **Additional qualitative examples and failure modes.** Underlined text marks failure modes discussed in Appendix E.

pendix quantifies both phenomena on the MIMIC-CXR test set.

F.1 Prompt-Decoder Agreement

To assess the fidelity of the decoder to the diagnostic prompt, we extract CheXbert labels from each generated report and compare, for every (image, condition) pair, the state predicted by the prototype head with the state expressed in the generated text. Agreement is observed in 96.7% of all pairs. The clinically relevant discrepancy, where the decoder introduces a positive finding that the prototype head did not predict as such, accounts for 1.1% of pairs.

To further characterise the influence of the diagnostic prompt on the generated output, we compare reports produced under the standard inference regime (predicted prompts) with reports produced under an oracle regime in which ground-truth states replace the classifier output, mirroring the teacher-forcing setting used during training.

As shown in Table 10, oracle prompts yield an absolute F1 improvement of 0.31, confirming that the clinical content of the generated reports is substantially determined by the diagnostic prompt. This result indicates that a considerable fraction of the error observed under standard inference is attributable to inaccuracies in the prototype-head predictions rather than to the decoder disregarding the prompt.

Table 10: Clinical metrics under predicted and ground-truth diagnostic prompts on the MIMIC-CXR test set. Both configurations correspond to a single training run.

Prompt source	P	R	F1
Predicted (standard inference)	0.487	0.556	0.490
Ground-truth (oracle)	0.810	0.825	0.803

F.2 Provenance of False Positive Findings

We analyse the provenance of false positive findings by identifying conditions that are expressed as positive in the generated report but labelled as negative or blank in the reference. For each such instance, we determine whether the same condition appears in any of the $K=3$ reports retrieved for that image. In 30.4% of cases, the false positive condition is absent from all retrieved reports, ruling out retrieval as a direct source. The remaining 69.6%, in which the condition appears in at least one retrieved report, are retained as candidates for

textual propagation analysis; co-occurrence alone does not establish a causal link.

To distinguish retrieval-induced overlap from overlap attributable to the formulaic nature of radiology language, we measure, for each candidate case i, the length $L_{\text{ret},i}$ of the longest contiguous token sequence shared between the generated false positive sentence and the corresponding sentences in the retrieved reports. A matched control quantity $L_{\text{ctrl},i}$ is computed analogously using training reports that express the same condition but were not provided to the decoder.

The paired margin $m_i = L_{\text{ret},i} - L_{\text{ctrl},i}$ captures the excess overlap attributable to retrieval. The median overlap is two tokens in both groups, and the mean margin is -0.24 tokens (95% CI: $[-0.29, -0.18]$), indicating that the retrieved reports exhibit no greater verbatim similarity than the matched controls. The two-token median is consistent with the recurrence of standard clinical phrases across radiology reports.

To account for potential outliers, we examine the high-overlap tail defined by $m_i \geq 5$. Only 0.7% of candidate cases meet this criterion. These instances are retained as conservative candidates for direct textual propagation; however, they represent a negligible fraction of the overall false positive population.